中国語学習者を対象にしたスピーキングテストの評価尺度の開発

著者	曲 明
雑誌名	外国語教育研究
巻	19
ページ	39-57
発行年	2016
URL	http://hdl.handle.net/10258/00009550

『外国語教育研究』外国語教育学会紀要 No.19 (2016) pp.39-57.

中国語学習者を対象にしたスピーキングテストの評価尺度の開発

曲 明

1. はじめに

外国語教育の評価に関する議論の中で、目標基準準拠評価(criterion-referenced assesment)、パフォーマンス評価が注目されるようになった。中国語教育分野においても、スピーキング能力を養成するための、より良いスピーキング能力の評価方法が求められている。しかし、中国語教育分野においでは、教育目標を「コミュニケーション能力の向上」と定めているにもかかわらず、スピーキング能力の評価尺度はいまだ確立されていない。現場では、スピーキングクラスにおける期末テストの実施の際、スピーキングテストの代わりにペーパーテストを用いる教員もいれば、勘や印象によって学生のスピーキング能力を評価する教員もいる。このような状況においては、信頼性、妥当性、実用性が高いスピーキング能力の評価方法が必要不可欠である。本研究では、中国語スピーキング能力の熟達度を評価する際に教育現場で使用しやすい評価尺度の開発を目指す。

2. 理論的考察

2.1 スピーキングテストの種類

Luoma (2004) はスピーキングテストを以下のように分類している。①間接測定法 (indirect measurement)、②半直接測定法 (semi-direct measurement)、③直接測定法 (direct measurement)である。

「間接測定法」とは、他のテスト形式(ペーパーテスト、クローズテストなど)によってスピーキング力を測ろうとするものである。テストとしての妥当性、波及効果などの視点から、一般的に間接測定法については批判する論が多い。「半直接測定法」は、録音された受験者の発話を評価する方法である。会話の相手とのやりとりがなく、機械を利用して受験者の音声を録音し、評価する。このタイプのテストは現実社会での言語使用状況から少し離れているが、テスト時間の節約になることや、実施条件が統一しやすいことから、テストの信頼性が高いとされている。「直接測定法」は、対話者と

のやりとりの中から発話が引き出されていくタイプのテストである。具体例として、インタビューテスト、ペアテスト、グループテストなどが挙げられる。

インタビューテストは、スピーキングテストの中で最も行われている方法である。しかし、コミュニケーションの方向は一方的で、質問者と受験者の関係が非対称的である。会話の場面において求められる自発的な発信能力や会話管理能力が十分に評価されていない可能性があると言われている(Young, 1995;Fulcher,1996 など)。教育現場の大規模教室において用いる場合、インタビューテストでは時間や経費がかかりすぎるというテストの実用性に関する問題もある(Young, 1995)。ペアテストは二人の非母語話者の受験者を同時に面接する方法である。もっとも有名なペアテストは、イギリスのケンブリッジ大学海外試験評議会の英語検定試験である。ペアテストは受験者のストレスを軽減し(Ikeda, 1998)、インタビューテストに比べて双方向の釣り合いのとれた発話が生まれる(May, 2002)とされている。ペアテストの問題点として、受験者の発話がペアを組む相手の発話に影響される可能性があると指摘されている(Hughes, 2003)。グループテストは 3 人以上の受験者を同時に評価するテスト形式である。テストの表面妥当性が高く、スピーキングテストの形式の 1 つとして、受験者にも好評であると報告されている (Shohamy et al., 1986; Fulcher, 1996)。しかしグループメンバーの各々の言語レベル、性格などの個人的な要因が、ペアテスト以上にテスト結果に影響を及ぼすため、利害関係の高い(High stakes)テストにはあまり使われていない。

近年、外国語教育現場では、「実践的コミュニケーション能力を育てる」という教育方針に基づき、授業中にペアワークを行う教育現場が多いと思われる。授業と評価の一貫性を追求するという意味からも、本研究では教育現場において実用性が高いペアテストの評価尺度を開発したい。

2.2 スピーキングテストの評価尺度の種類

スピーキング能力の評価基準の設定には2つのタイプが認められる。1つは分析的評価(analytic evaluation)で、もう1つは全体的評価(holistic evaluation)である。分析的評価は、文法、語彙、発音、流暢さ、内容などの評価項目に分けて評価を行う方法である。全体的評価とは個別的な項目には着目せずに、全体として 1 つの印象点を出す方法である。

分析的評価は、全体的評価より高い信頼性が得やすい。学習者へのフィードバックも診断的な機能を持たせやすい。一方、問題点として、スピーキングパフォーマ

ンスを仮に文法、語彙、発音などの下位項目ごとに評価しようとしても、全体的な印象に影響されて個別の評価作業が曖昧になってしまう可能性がある(馬場, 1997;根岸, 2005)ことが挙げられる。全体的評価は採点者にとって楽な方法であるが、しかし、各下位項目の中でレベルに差があったとしても、すべて単一の全体的評価しか表されないため、どのような点を改めれば、その人のパフォーマンスが向上するかという診断的情報が得られないと指摘されている(馬場, 1997)。

どの採点方法にもそれぞれ長所と短所がある。本研究では、学習者へのフィードバック、教育効果及びテストの波及効果を考えつつ、分析的評価の尺度を開発したいと考える。

2.3 評価尺度を開発するアプローチ

評価尺度の開発に当たっては、大別すると2種類がある。既存の尺度の能力記述文からのアプローチと、言語行動の例からのアプローチがある(Council of Europe, 2004; 猫田, 2007)。前者においては、まず「何について記述するのか」という記述の対象をあらかじめ決定した上で、既存の記述子(descriptors)から自分のテストに相応しい記述子を収集することで評価尺度を作成する。このアプローチの長所は、様々な既存の理論的枠組みを利用することで、一般化を試みることが容認されやすい点にある。短所としては、事後的な検証をしても、一度作成された記述子は人々に無条件に受け入れられてしまう危険がある(猫田, 2007)ことが挙げられる。後者のアプローチにおいては、学習者の言語行動の標本を実際に観察、評価することによって、評価者がどのような言語的特徴に注目し、どのような判断を下すかを探る。そして、そのデータに基づいて記述子を帰納的に言語化、尺度化にする。後者の短所は、最終的にどのような尺度が完成するかがデータに大きく依存してしまうことである。(猫田, 2007)。本研究は日本の中国語教育現場で使用する評価尺度を考案する、はじめての試みとして、なるべく多数の教育現場に貢献したい。また、一般化しやすいものを作成すべきと考えるため、既存の評価尺度と記述子に基づいた前者のアプローチによって評価尺度の開発を行う。

2.4 評価尺度の妥当性の検証

Council of Europe(2004)によれば、既存の評価尺度の能力記述文から評価尺度を開発する際に、おおよそ以下のプロセスを辿る。教育理論及び言語学理論に基づいて評価項目、評価ポイントを決定したうえで、既存の評価尺度から、教員にとっ

てわかり易いと判断された記述子を選び、暫定的な評価尺度を作成する。その後、作成された評価尺度の妥当性を検証する。評価尺度の妥当性を検証するには、質的な方法と量的な方法がある。質的な方法は、現職の教員が作成された評価尺度を用いて学生のスピーキングパフォーマンスを採点し、記述子の解釈しやすさ、評価尺度の使いやすさについて評価して、その後、教員たちの評価に基づいて、記述子の質的検証、修正、追加作成を行うことである(Brown,2000;2003; 猫田, 2007)。近年では、スピーキングテストの評価尺度を質的に検証するものが増えてきており、英語スピーキングテストの評価尺度の妥当性の検証を行ったのは、Brown(2000)、Brown(2003)、Iwashita & McNamara (2005) などがある。量的な方法による評価尺度の妥当性の検証は、実際のパフォーマンス評価データを多相ラッシュ・モデル(FACETS)を用いて分析することにより、各評価項目の困難度、モデルへの適合度を算出することである。これらの方法は、ヨーロッパ教育協議会が各種の共通参照尺度を開発する際に用いた方法に沿ったものである(North,2000; Council of Europe,2004)。 なお、本研究において、ペアテストの評価尺度を開発するにあたっては、本来であれば、上述の両方の方法による妥当性を検証するのが理想的ではあるのだが、紙幅の制限上、本稿では質的な方法による検証にとどめ、量的な方法による検証については、稿を改めることとする。

3. 研究方法

本研究の目的は中国語スピーキング能力の熟達度を評価する際に教育現場で使用しやすいペアテストの評価尺度を開発することである。開発の作業は、Council of Europe(2004)の提起する評価尺度開発に関わる理論に従い、3 つの段階を踏んで進めてゆく。①日本国内外の中国語スピーキングテストの評価尺度をレビューし、日本語母語話者の中国語学習者を対象としたスピーキングテストの形式に相応しい評価尺度の下位項目を決める。②①の段階で決めた評価項目と日本における中国語教育の到達目標を照らし合わせながら、各評価項目の評価ポイントを決め、先行研究から記述子を選び、暫定的な評価尺度を作る。③暫定的な評価尺度を用い、20 人の中国語スピーキングテストのデータを 5 人の教員に評価してもらい、アンケート調査によって評価尺度の妥当性を検証する。

3.1 国内外の中国語スピーキングテストの評価尺度

近年、日本国内外において、利害関係が高い中国語スピーキングテストがいくつか開発された。中国政府及び大学に開発されたものには HSKK、BCT、ATSC、CST などがある。日本国内で開発されたものには中国語検定試験二次試験がある。BCT はビジネス中国語のテストで、社会人を対象に開発されたものであり、本研究の研究対象とは異なる。CST はパソコンによるテストであるが、全体的評価尺度を用いているため、本研究で扱う分析的評価尺度と異なる。この二つのテストの詳細について、紙幅の関係でここでは特に述べない。この節では、BCT、CST 以外の各試験の概要、用いる評価尺度について簡単に紹介し、本研究で用いるテストに相応しい評価項目及び評価観点を定めたい。

・HSKK（Hànyǔ Shuǐpíng Kǒushì Kǎoshì）

HSK は、中国教育部国家漢語弁公室が主催し、漢語を非母語とする者の中国語水準を測るための国家試験である。HSKK は HSK のスピーキングパートで、受験生のスピーキング能力を判定するテストである。初、中、高級の 3 つのレベルがある。受験者の発話が録音され、その録音に基づいて評価される半直接テストである。HSKK の試験内容は 3 つの部分からなる。①復唱：放送を聴いて、その文書を復唱する。②聞き取り：質問を聞いて、それについて簡潔に答える。絵を見て話す：問題用紙にある 1 枚の絵を見ながら、それについて話す。③読み取り、質問に答える：問題用紙に書かれた質問に対して答える。評価項目と評価観点は以下に示す通りである。文法(文型の多様性、正確さ)、語彙(正確さ、豊富さ)、話題の展開（首尾一貫したまとまった内容が言えるかどうか）、発音(聞き取れるかどうか、ピンイン、語調の把握)である(下線は筆者による、以下同様)。

・中国語検定試験二次試験（日本）

中国語検定試験は日本中国語検定協会により開発されたテストである。その準一級、一級では、一次試験合格者を対象に、二次試験でスピーキングのテストが課される。試験で求められる能力は、全体的に社会生活に必要な中国語を習得し，通常の文章の中国語訳・日本語訳、また簡単な通訳ができることである。通訳する際、単に単語を知っているということから一歩先に進んで、熟語や慣用句にも精通していることに加えて、時事問題にも対応できる力が求められる。試験のタスクタイプは以下の 3 つがあり、①質問に答える、②翻訳する、③あるテーマについて自分の意見を述べるである。また、ホームページ上に公開された受験案内によれば、スピーキングテストの評価項目は以下のようになる、「語彙と文法を運用する力および発音の状況、中国語によるコミュニケーション能力、翻訳力」である。

・ATSC(Automated Test of Spoken Chinese)

ATSC は中国語版 Versant と呼ばれ、近年、北京大学とアメリカのテスト開発会社によって共同開発された電話によるスピーキングテストである。被験者は電話でテストを受け、テスト終了 2 分後に評価結果が出るという半直接テストである。問題のタスクタイプは以下の 5 つがあり、①朗读句子(文を朗読する)、②重复句子(聞いた文を復唱する)、③回答问题（聞かれた質問に答える)、④重组句子(提示された単語やフレーズを並べ替えて、口頭で文を作る)、⑤短文重述(読んだ文の意味を自分の言葉で再現する）である。評価尺度は以下の4つで、句子的掌握(文の正しさ）、词汇(語彙)、流利度(流暢さ)、发音、声调(発音、声調)である。

上述 3 つの中国語スピーキングテストはすべて、モノローグタスクと「質問に答える」というコミュニケーションの方向が一方的であるタスクを使用している。本研究で扱うペアテストで用いる受験者が双方に質問し合うタスクと異なり、評価尺度を決める際に日本で教育現場のために開発された英語のグループスピーキングテストの KEPTも参考にしたい。KEPT の概要は以下に述べる。

・KEPT(Kanda English Proficiency Test)

KEPT は神田外語大学で開発されたインタアクティブタスクを使った英語スピーキングテストである。KEPT はグループテストという形式をとり、3 人の学習者同士の話合いを評価する。評価尺度には発音、流暢さ、語彙、文法のほか、コミュニケーションスキルが付け加えられている。KEPT は唯一日本の教育現場のためだけに開発された、利害関係が低い(Lowstakes)テストであると言われている(Bonk & Ockey,2003)。また、KEPT が開発されたあと、その評価尺度の妥当性が検証されており、高い妥当性を持っていることが証明されている(Bonk & Ockey,2003)。

上記の 4 つのテストの評価項目には発音、語彙、文法が含まれている。このことから、この 3 つの評価項目が、スピーキング能力を構成する要素として先行研究で認められていることが示された。ペアテストで用いるタスクはインタラクティブタスクであり、評価尺度を決める際、上記 3 つの評価項目以外、「流暢さ」と「コミュニケーションスキル」の 2 項目を評価尺度に付け加えたい。その理由は以下に述べる。まず、「流暢さ」は多くのスピーキングテスト(ATSC, KEPT, VERSANT など)において評価項目として取り入れられていることが挙げられる。また、会話において、ある言語機能や目的を果たそうとする際に、成功するか否かを決定する要因の1つは「流暢さ」である(Council of Europe, 2004)と言われているため、スピーキング能力を評価する際には欠かせない側面であると言えよう。また、本研究で扱うテスト形式は対面式ペアテス

トであるため、会話の相手とのコミュニケーションを効果的に行うことができたかどうかも、本研究のテストで測りたいスピーキング能力の側面である。以上の点を踏まえて、KEPT と同様、「コミュニケーションスキル」も 1 つの評価項目として選ぶこととする。

3.2 暫定的な評価尺度の作成

評価項目を定めるにあたっては、日本国内外で開発されたスピーキングテストの評価尺度以外に、日本の各大学の中国語スピーキングクラスの教育目標も参考にした。各大学のホームページに掲載されている中国語スピーキング授業のシラバスによれば、多くの大学は「中国語によるコミュニケーション能力の向上」を教育目標としている。しかし、「コミュニケーション能力」を向上させるためには、どのような下位能力を発達させなければならないかについては、ほとんどのシラバスでは言及されていない。中国語スピーキング能力の教育目標を具体的に記述したものに、《国际汉语教学通用课程大纲》(以下《大纲》と略す)がある。《大纲》は 2008 年に発布された、外国人への中国語教育用に制定した標準教育課程である。中国における対外中国語教育をはじめ、世界各地の中国語教育のスタンダードとして示されたものである。《大纲》の示す中国語スピーキング能力の教育目標の中には、発音(ピンインの把握、アクセントの自然さ)、正確さ(語彙、文法の正確さ)、豊富さ(幅広い語彙、文法項目を使えるか)、流暢さ(理解に妨げる不自然なポーズがあるか)、コミュニケーションストラテジーの使用が含まれており、それぞれの項目について詳細に記述されている。ここで挙げられている 5 つの項目は前節で紹介した各テストで用いている評価項目と重なっており、それを踏まえて、本研究で扱うペアテストの評価項目も上記の 5 つの項目、すなわち、発音、語彙、文法、流暢さ、コミュニケーションスキルと定めた。上記それぞれの項目がどのような能力を指しているのか、つまり記述子を選ぶ際に、どのようなものを選ぶべきなのかを判断するために、それぞれの評価項目で評価するポイントを中国語教育学理論に基づき決めた。

◆音声能力(発音)

輿水(2005)によれば、中国語の音声知識は以下のものを含む。

声調:4 種類の声調に関する知識。
声調の組み合わせ:声調の組み合わせによる声調の変化(変調)に関する知識
音節の構成:音節＝声母＋(韻母、韻頭/韻腹/韻尾)
韻母:単母音、二重母音、三重母音、尾音、捲舌母音の発音に関する知識
声母:有気音、無気音、両唇音,唇歯音,舌尖音,舌根音,舌面音,舌歯音に関す

る発音の知識

本研究では音声能力（本研究では発音と称するが）について、上述の音声知識を認識し、発声できることと捉える。しかし、発音を評価する際、瞬間的に声母、韻母など細かい部分を弁別できないと考え、発音の評価は主に自然かどうか、理解可能かどうかに焦点を当てる。評価ポイントは、以下の 2 つに定めた。

・日本人母語話者に特有の発音の問題が多くあるのか。

・個々の音が曖昧ではなく、聞き手が容易に理解できる明瞭さで発音されているか。

◆語彙能力

本研究では、言語の語彙能力について、語彙知識を有し、更にその語彙を使いこなす能力であると捉える。語彙能力の評価尺度を作る際には語彙知識の豊富さ（語彙の幅、レパートリー）と語彙を使いこなす能力（語彙の適切さ）の 2 つの評価ポイントを定めた。

・定型表現や口語表現を含めた様々な語彙使用の幅

・語彙使用の適切さ

◆　文法能力

本研究においては、「文法」という言葉を以下のように捉える。すなわち、外国語として中国語を教えたり学んだりすることを目的とし、教育的な配慮のもとに便宜的に文法のルールをまとめた、「中国語教育文法」のことを指すものとする。北京語言大学出版社から刊行された《对外汉语教学实用语法》（卢福波，2010）では、文法は“词”の部分と“句子”の部分に分類されている。前者においては、品詞について、名詞、代詞、動詞、形容詞、数詞、量詞、前置詞、副詞、連詞、助詞、間投詞の順に説明されており、後者においては文の分類、主語・述語・目的語、及びそれ以外の文成分、時間とアスペクト、種種の述語文、否定、強調、複文と 7 項目に分けて説明がなされ、文法項目として 170 個が盛り込まれている。紙幅の関係から、ここでは詳述しないが、本研究ではこれらの知識を文法知識とし、それを使いこなせる能力を文法能力と捉え、以下のように文法項目の評価ポイントを定めた。

・日常的な話題を扱ったり、意見を述べたりする時、妨げになるような文法上の間違いはないか。

・受身文、使役文、存現文、“把”構文、兼語文など広い範囲の多種多様な文法項目を扱えるか。

◆流暢さ

流暢さを客観的に示す言語指標には、大別して、会話の一時的な停止を捉え

た場合と、言い淀みを捉えた場合の 2 つに分類される(Foster & Skehan, 1996)。前者については、不自然なポーズの数を数える方法やポーズの時間を測定する方法が考案されている。後者については非流暢さの尺度として、繰り返し、出だしの言い間違い、自己訂正などを数える方法がある(堀川, 2007)。本研究では、上記の両側面の言語指標に注目しつつ、先行研究で用いている評価尺度を参考に、流暢さの評価ポイントを次のように定めた。

・自然で安定したスピードで日常の話題を話せるか。

・言い換えたり、語彙、表現を探したりするときに、理解の妨げになる程の長いポーズがあるか、発話がたどたどしくなったりせずに話せるか。

◆コミュニケーションスキル

コミュニケーションスキルに関しては、KEPT で用いる指標に従い、「コミュニケーションストラテジーを用いて、効果的にコミュニケーションを行うスキル」と捉える。コミュニケーションストラテジーの定義は、問題解決という限られた場面における方略から、言語学習または認知という大きな枠組みの中で捉えた広い意味でのストラテジーまで分類されている。しかし、それらの分類は、言語活動の中で、何らかの困難に遭遇した場合、コミュニケーションを続けるためにとる方略を含んでいるという点では、共通する部分も多い(岩井, 2000)。そこで本研究では、「学習者が、発話を進めて行く上で、語彙・文法などの知識不足により、理解・産出に困難を感じた場合、発話を維持するためにとる方略」をコミュニケーションストラテジーの定義とする。具体的には、以下のようにコミュニケーションスキルの評価ポイントを決めた。

・コミュニケーションの挫折に遭遇するときに多様なコミュニケーションストラテジーを効果的に使用し、話を続けることができるか。

・相手と共同で話を作り、展開できるか。自分の発話を相手の発話に上手に関連付けて会話することができるか。

評定項目と評価観点を決定した後、各評価項目のレベル別の記述を行った。各評価項目の記述子は主に CEFR のスピーキング能力に関する記述、HSKK、KEPT の評価尺度(日本語訳は堀川(2003)を参照した)、《大纲》が示している中国語の教育目標および先行研究で用いられているテストの評価尺度をもとに選んだ。また、先行研究の評価尺度の習慣に従い、5 段階評価法を採用した。草案作成にあたり、4 名の中国語教師の意見を聞いて調整を行った。作成した評価尺度については、付録1を参照されたい。

4　暫定的な評価尺度の妥当性の検証

4.1　妥当性の検証方法

本研究では、アンケート調査という方法により、評価尺度の妥当性の検証を行う。日本の大学で中国語を専攻する大学生 20 人(詳細は表 2 を参照されたい)にペアテストを受けてもらい、テストにおける発話をすべて録音した。中国語の教員 5 人(日本語母語話者 3 人、中国語母語話者 2 人、ともに教育経験は 10 年以上である)にそれぞれ 4 人(2 ペア)のデータを暫定的な評価尺度を用いて採点してもらった。得られた成績の詳細を表 3 に示す。なお、今回の採点に当たり、事前に採点者たちに評価尺度、評価手順の説明をしたあと、1 ペアのテストパフォーマンスを聞きながら、1 時間ほどの採点者トレーニングを実施した。採点終了後、採点者たちに対して、評価尺度の内容をどのように理解して採点したのか、評価尺度は使いやすいか、各評価項目の 5 段階レベルは区別しやすいかについて、あらかじめ用意しておいたアンケート用紙を用いて、アンケート調査を実施した。アンケート調査の質問の内容は Brown (2003)を参考にした。

表 2　テスト参加者情報

	評価者	学生	性別	学年	成績				
					P	V	G	F	C
ペア 1	A	1	F	2	3	3	3	3	2
		2	F	2	3	3	3	2	2
ペア 2		3	F	2	2	3	3	3	3
		4	F	2	3	3	3	2	2
ペア 3	B	5	F	2	3	3	3	3	3
		6	F	2	3	2	3	3	2
ペア 4		7	F	2	3	3	2	3	2
		8	M	2	4	3	3	3	3
ペア 5	C	9	F	2	3	3	3	3	3
		10	F	2	3	3	3	2	2
ペア 6		11	F	2	2	2	3	2	2

		12	F	2	3	2	3	2	3
ペア 7	D	13	M	3	4	4	4	4	3
		14	F	3	4	4	4	3	3
ペア 8		15	F	3	3	4	4	4	3
		16	F	3	3	4	4	3	3
ペア 9	E	17	M	3	4	4	4	4	4
		18	F	3	3	4	4	4	3
ペ ア 10		19	F	3	3	4	4	4	3
		20	F	3	3	3	3	3	3

注:性別の欄にあるFは女子学生、Mは男子学生である。成績の欄にある P は発音、V は語彙、G は文法、F は流暢さ、C はコミュニケーションスキルである。

表 3 各評価項目の平均値と標準偏差

評価尺度	平均(M)	標準偏差(SD)
発音	3.1	.53
語彙	3.2	.67
文法	3.3	.55
流暢さ	3	.70
コミュニケーションスキル	2.7	.55

表 3 によれば、今回のテストデータにおいて、もっとも平均値が高いのは文法の項目であり(m＝3.3)、もっとも平均値が低いのはコミュニケーションスキル(m=2.7)である。この結果から、被験者たちは文法的に正しい発話をし、コミュニケーションストラテジーの使用はあまり活発ではないことがわかった。また、標準偏差の値に関しては、語彙と流暢さの値は高いが、発音、文法、コミュニケーションスキルの値は低かった。この結果から、語彙と流暢さの点数はばらつきが大きく、発音、文法、コミュニケーションスキルの点数はばらつきが小さいことがわかった。

4.2 アンケート調査の結果

アンケート調査の質問は 9 個からなる。この節では質問の順にアンケート調査の結果を述べる。

① 「ペアテストの会話能力を測るのに、この5つの評価項目で十分でしたか？　他に立てるべき項目はありますか？」
採点者5人のうち、1人がこの5つの評価項目で「問題がない」と答えた。残りの4人は「話の内容(面白い話であるかどうか、ユーモアが混じって話せるか否か)、態度(積極的に話すか、しないか)、発話量(たくさん話すかどうか)も評価の対象にした方が良いのではないか」と答えた。その理由として、自分たちの評価が上記のポイントに影響されやすいものであるということが挙げられた。

② 「この評価尺度はすべての学生の能力をカバーしていますか？」
採点者5人全員が「すべての学生の能力をカバーしている」と答えた。

③ 「教育現場の先生にとって、この評価尺度は使いやすいですか？」
採点者5人のうち、3人が「まあまあ使いやすい」と答えた。残りの2人は「採点者トレーニングを受けないとあまり自信がない」、「重要な記述子の用語について詳細な説明をしてほしい」と答えた。

④ 「この評価尺度に分かりにくい用語、あるいは不適切な記述はありましたか？」
採点者5人のうち、4人は「特に問題がない」と答えたが、残りの1人は「コミュニケーションスキルの概念を詳細に記述して欲しい」と答えた。

⑤ 「すべての記述子は理解しやすく、解釈可能でしたか？自分の評価にどれぐらい自信を持っていますか？下の表に自己採点をし、その理由も教えてください。」
採点者5人のうち、3人が「記述子はおおよそ理解しやすく、解釈可能」と答えた。1人が「コミュニケーションスキルへの解釈に自信がない」と答えた。自分の評価にどれぐらい自信を持っているかを聞いたところ、表4のような結果になった。

表4　自分の評価にどれぐらい自信を持つか

	採点者					
	1	2	3	4	5	Mean
発音	5	4	4	5	5	4.6
語彙	4	5	4	3	4	4
文法	5	5	4	4	5	4.6
流暢さ	5	4	4	4	4	4.2
コミュニケーション	4	4	4	3	3	3.6

表4によれば、「自信がない」を1点、「とても自信がある」を5点にしたところ、

コミュニケーション項目以外の項目の平均値はすべて 4 点を超えた。コミュニケーション項目の平均値は 3.6 点となった。この結果から、コミュニケーション項目以外の 4 項目の採点に対して、採点者たちが自分の評価に自信を持っていることが分かった。その理由について、項目ごとに下記のコメントが得られた。

・発音:3 人の採点者が「とても自信がある」の 5 点を選び、残りの 2 人は 4 点を選んだ。全員に「自信がある」という結果となった。5 点を選んだ採点者は、「発音の善し悪しはわかりやすい、日本語母語話者の発音の特徴が目立つ」ということを理由として挙げた。4 点を選んだ中国語母語話者の採点者は「基本的に通じれば良いという感覚を持っているので、聞き逃したポイントがあるかもしれない。ほかの項目ほど真剣ではなかった」とコメントを残した。

・語彙:4〜5 点を選んだ人は 3 人で、1 人が 3 点を選んだ。「自信がある」と答えた人は、「四字熟語、ことわざ、定型表現がヒントになる」とコメントを残していた。

・文法:全員 4〜5 点を選び、全員「自信がある」という結果となった。その理由として、「文法の間違いはわかりやすい」、「文型が単調であることは目立つ」という点が挙げられた。

・流暢さ:5 点を選ぶ人は 1 人で、ほかの 4 人は 4 点を選んだ。「自信を持っている理由」として、「ポーズが自然であるか否かはわかりやすい」、「性格によってポーズをとったり、話すスピードが遅かったりする学生もいるが、それもなんとなく分かる」といったことが挙げられた。

・コミュニケーションスキル:評価者 5 人のうち、3 人が 4 点を選び、2 人が 3 点を選んだ。「自信を持つ」得点の平均値は 5 項目の中で最も低かった。その理由として、「自分の発話は相手の発話に関連つけることが出来るかどうかは判断しやすいが、コミュニケーションストラテジーは使用されないことがあるため、判断しにくい場合がある」、「言語習熟度が低い受験者のディスカッションは活発ではないので、自分の発話を相手の発話に関連つけなくてもタスクの完成がなんとなく出来る」という点が挙げられた。

⑥「各項目で評価している内容は重複(overlap)している部分がありますか？」

評価内容が重複していると思われた項目は 2 つあった。

語彙と文法:「複文が使われる際に、接続詞が間違っていると、結局語彙も、文法も減点することになる。評価している内容は重複していると思う」との指摘があった。

流暢さと語彙:「接続詞及び談話標識(discourse markers)の使用に対する評

価と、流暢さに対する評価が重複しているのではないか」とのコメントが得られた。

⑦「評価項目の記述文はもっと長い(短い)方が良かったですか？」

記述文の長さについて聞いたところ、採点者全員が「これで良い」と答えた。

⑧「各評価項目の点数をつける際に、迷いはありましたか？5 段階レベルの区別は判断しやすかったですか？」

「全体として、5 点はあげにくい」というコメントが得られた。その理由として「5 点の記述に「いつもに正確、一貫して正しく、全般的に適切」といったことが書かれており、なかなかそのような学生がいない」という点が挙げられた。同様に、「1 点の採点も少ない、そのような学生も少ないから」という意見が 2 人の採点者から得られた。

⑨「その他何か気付いたことがありましたら、気軽に書いてください。」

その他を聞いたところ、「流暢さはトピックの熟知度(familiarity)にも影響されると思う」というコメントが得られた。

5.考察

アンケート調査 1 番の質問に対する、「話の内容、ディスカッションへの態度、発話量も評価項目に入れるべきではないか」という指摘については、「話の内容」の良し悪しは採点者の主観的な判断に委ねることが多いため、今回の評価項目には含めないものとする。「発話量」および「積極的な態度」に極端に差がある場合、試験官が介入することにより調整可能と考え、「内容」と同様、評価項目に入れないことにする。上記の 3 つのポイントに影響されやすいということは想像に難くないが、それを改善するには、今後における採点者トレーニングを実施する際に更なる説明が必要と思われる。

「コミュニケーションスキル」の採点について、「自信がある」の項目の得点が低かったことの理由は、書かれたコメントの中にもあったが、コミュニケーションスキルの概念に詳しくない教員もいるためだと思われる。インタラクティブタスクを使ったスピーキングテストは今まで開発されてこなかったこともあり、会話の管理能力、コミュニケーションストラテジーについては、やはり今後採点者トレーニングを実施する際、評価ポイントについて、明示的に記すべきだと思われる。

評価する内容が重複していると思われた接続詞の使用と複文の使用に関して、複文中に接続詞が使用されるとき、その接続詞は「文法」の項目として捕らえるべきと

考える。また、学生の発話の中で、会話をスムーズにするためにほとんど意識をせずに発する談話標識(例えば、中国語の“那”、“还有”など)の使用は、本来ならば「語彙」の項目において採点されるべきだが、しかし、その適切な使用は会話の流れを良くし、流暢さにも貢献するため、談話標識に対する採点は、確かに語彙と流暢さの両方に跨っているものと思われる。スピーキングテストの評価尺度について、「流暢さ」の得点と「語彙項目」の得点の相関性が高いことを指摘する研究もある(Chapelle et al., 2008)。今後、この両項目の評価尺度については、さらに調整する必要があると思われる。

最高点の 5 点と最低点の 1 点は選びにくいという点については、採点者の指摘にあったように、記述文において、「すべて、一貫して、全般的に」といった言葉が使用されていることが理由だと思われ、記述文は再検討する必要がある。その際、1 点の記述文の文案は、言語習熟度が低い学生に対して使用できるかどうかを検討したうえで決める。

最後に、評価尺度の妥当性の検証とは直接には関係ないが、「流暢さはトピックの熟知度に影響される」、「タスクをもっと収束的なもの(convergent)にしないと、コミュニケーションストラテジーがあまり使用されない」といった貴重な指摘も得られた。これらの意見をペアテストのタスク開発に生かしたい。

6. おわりに

本研究では、一般教育現場で用いる中国語ペアテストの評価尺度を開発した。テスト理論、日本国内外の評価尺度、および中国語教育の目標について考察した上で、暫定的な評価尺度を定めた。この暫定的な評価尺度を用いて、日本語母語話者の中国語ペアテストのデータを 5 人の中国語教員に評価してもらい、アンケート調査という方法により評価尺度の妥当性を検証した。多くの教員はこの評価尺度は使いやすく、記述子の解釈もしやすいと評価した。各評価項目間の重複および 5 段階のレベル分けのしやすさについて、建設的な意見も得られた。今後評価尺度を改善する必要があり、採点者トレーニングのやり方についても考え直す必要がある。

スピーキングテストを実施する際、テストの評価基準を学生に明確に示すことによって、学生は今の自分がどれぐらいの能力を持っているのか、次へのステップアップを目指すためにはどのような能力を身につけなければいけないのかを明確に認識することができる。さらに、受験者はテストを受ける前からテストの評価基準を推察し、そ

れに合わせて受験準備を取り組むことができれば、試験を行う人が望むテストの波及効果も得られる。

本研究では、データの数が少なく、受験者の能力の幅が狭いため、結果の一般化には限界がある。今後広い能力幅の学習者からデータを取り、採点者の人数も増やした大規模の調査が望まれる。また、先に述べた通り、量的な方法による妥当性の検証についても、今後の課題としたい。

（室蘭工業大学）

参考文献

欧文

Bonk,J. & Ockey, J. (2003). A many-facet Rasch analysis of the second language group oral discussion task, *Language Testing*, 20(1).

Brown,A. (2000). An investigation of the rating process in the IELTS Speaking Module, *IELTS Research reporta*, vo13.

Brown,A. (2003). An examination of the rating process in the revised IELTS Speaking Test, *IELTS Research reporta*,vo6.

Brown,A. (2005). *Interviewer variability in oral proficiency interviews*, Peter Lang, Frankfurt.

Brown,A, N Iwashita, T McNamara. (2005). An examination of rater orientations and test-taker performance on English for academic purpose speaking tasks, *TOEFL Monograph series MS-29*, Educational Testing Service, Princeton, New Jersey.

Chapelle et al.(2008). *Building a Validity Argument for the Test of English as a Foreign Language*, NewYork: Routledge.

Hughes, A. (2003). *Testing for Language Teachers*, Cambridge : Cambridge University Press.

Ikeda, K. (1998). The Paired Learner Interview: A Preliminary Investigation Applying Vygotskian Insights, *Language, Culture and Curriculum*, 11(1).

Luoma, S. (2004). *Assessing Speaking.* Cambridge: Cambridge University Press.

May, L. (2002). An Exploration of the Validity of the Paired Candidate interaction, Abstract of the paper presented at 24th language Testing Research Colloquium.

North,B.(2000).*The development of a common framework scale of language proficiency,* NewYork: Peter Lang.

Foster, P.& Skehan, P. (1996). The influence of planning and task type on second Language performance, *Studies in Second Language Acquisition,* 18.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. Language Testing,13(1).

Saville, N. & Hargreaves. (1999). Assessing speaking in the revised FCE, *ELT Journal,* 53(1).

Shohamy, E, Reves T, & Bejarano Y. (1986). Introducing a new comprehensive test of oral proficiency, *ELT Journal,* 40(3).

Young,R.(1995). Conversational Styles in Language Proficiency Interviews, *Language Learning,* 45.

和文

岩井千秋(2000)『第二言語使用におけるコミュニケーション方略』, 渓水社.

Council of Europe (2004)『外国語の学習、教授、評価のためのヨーロッパ 共通参照枠』,(吉島 茂他訳), 朝日出版社.

輿水優(2005)『中国語の教え方・学び方−中国語科教育法概説』, 日本大学文理学部.

根岸雅史(2005)「スピーキングテストの採点はどうする」,『英語教育』, 第 7 巻.

猫田英伸(2007)「英語口頭運用技能の熟達度に関する記述子の開発」,『中国地区英語教育学会紀要』, 第 37 号.

馬場哲生(1997)『英語スピーキング論―話す力の育成と評価を科学する』, 河源社.

堀川有美(2003)「グループ・ディスカッション・テストにおけるテスト得点と受験者の発話データの分析」未発表修士論文 お茶の水大学

堀川有美 (2007)「日本語会話テストにおいてテスト形式が受験者の発話に与える影響」,『人間文化論厳』, お茶の水女子大学大学院人間文化研究科.

中国語

国家汉语国际推广领导小组办公室[編].(2008).『国际汉语教学通用课程大纲』,外语教学与研究出版社.

卢福波.2010.《对外汉语教学实用语法》北京语言大学出版社

付録1

暫定的な評価尺度

発音

1. 日本人母語話者に特有の発音の問題が多い。発話中にわからない発音があるため、意味が理解できないことが多い。
2. 日本人母語話者に特有の発音の問題がある。瞬間的にわからない発音があるが、文脈から理解できるものもある。意味理解を妨げることもある。
3. 発音ができていない箇所があるが、発話中に理解できることが多い。日本人母語話者に特有の発音の問題があるが、発話内容の理解を妨げない。
4. 発音の間違いが少々残っているが、意味理解の妨げにならない。
5. 発音が正しく、明瞭である。全体的にネイティブの発音に近い。

語彙

1. 語彙の幅が狭い。不適切な語彙・表現があり、意味をつかむことができない。
2. 語彙の幅にタスクを完成させるための広さがない、同じ言葉を繰り返し使用する。不適切な表現が多くあり、文脈からも意味を類推できない部分がある。
3. タスクを完成させるための語彙の幅を持っている。不適切な語彙・表現があるが、文脈から意味を理解できることが多い。
4. 幅広い語彙のレパートリーを使いこなせる。定型表現や口語表現も使える。たまに言い間違いがあるが、意味を理解するのに問題がない。
5. 定型表現や口語表現を含め、幅広い語彙を上手に使うことができる。一貫して正しく、適切に語彙を使用することができる。

文法

1. 発話の意図が類推できない文法の間違いがある。“是”を用いた表現が多く、使用する文法のパータンは極めて限定的、且つ単調である。
2. 理解を妨げる間違いがあるが、文脈から判断できるものもある。使用する構文の種類は少ないが、簡単な動詞述語文など基本的な文法表現が見られる。
3. 理解を妨げる間違いがあるが、殆ど文脈から判断できる。基本的な文法表現と簡単な複文(“因为”、“所以”、“虽然”、“但是”など)を使える。
4. 理解を妨げる文法の間違いがほぼない。文脈によって、“把”構文、兼語文など幅広い文法表現も見られる。複文が使える。

5. 理解を妨げる間違いがない。全般的に文法構造をよく理解しており、複文を含め、広い範囲の文法表現を適切に使える。

流暢さ

1. 発話に不自然なポーズが多い。聞き手を長く待たせることがある。日本人母語話者に特有の有声ポーズ(アー、エーなど)が多い。
2. 文を構築するために、言葉を選んだり、修正したりするためのポーズがあるが、話の要点は伝えられる。
3. 長めの発話で、ポーズがあったり、行き詰まったりすることはあるが、自力で話を続けることができる。
4. 難しいところでのやり直しや再構成があるが、自然体でスムーズに話を続けることができる。
5. 自然なリズムと安定したスピードで話せる。

コミュニケーションスキル

1. コミュニケーションのつまずきが多い。会話の相手から促しがなければ、最小限しか話さない。
2. コミュニケーションのつまずきが見られ、インタラクションが活発ではない。会話の相手からの促しがあれば、話をすこし展開する。
3. 話を展開するが、相手とのインタラクションに不適切な部分がある。基本的なコミュニケーションストラテジーが使えて、自分の発話を相手の発話に関連付けることができる。
4. 話を展開したり、続けたりすることが上手にできる。様々なコミュニケーションストラテジーを用い、自分の発言を他の話や相手の発言に関連付けることができる。
5. 効果的に各種コミュニケーションストラテジーを用い、相手とのインタラクションが活発にできる。前の発言に言及したり、示唆したりしながら会話することができる。

本稿は、科学研究費助成事業『中国語学習者を対象に発信力の向上を目指したスピーキングテストの開発』(基盤研究 C、課題番号 15K02781)による研究成果の一部である。